doi: https://doi.org/10.15446/rcp.v29n1.77214

Improving the Measurement of Children's Mental Health Problems in Colombia with Item Response Theory

JORGE CUARTAS

Harvard Graduate School of Education, Cambridge, United States

How to cite this article: Cuartas, J. (2020). Improving the Measurement of Children's Mental Health Problems in Colombia with Item Response Theory. *Revista Colombiana de Psicología, 29,* 87-103. https://doi.org/10.15446/rcp.v29n1.77214

Correspondence concerning this article should be addressed to Mr. Jorge Cuartas, e-mail: jcuartas@g.harvard.edu. Adress: Harvard University, 13 Appian Way, Cambridge, MA, The United States of America.

Acknowledgments: I would like to thank Andrew Ho for all his support in the process of writing this manuscript, and to Catalina Rey-Guerra, Ben West, and the members of S061 at the Harvard Graduate School of Education for their comments. All data for this study were taken from publicly available data sources. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

SCIENTIFIC RESEARCH ARTICLE

RECEIVED: JANUARY 12TH, 2019 – ACCEPTED: OCTOBER 19TH , 2019

Abstract

The present study examines the psychometric properties of the mental health scale for children used in the 2015 Colombian Mental Health Survey. To do so, a nationally representative sample of 2,727 children is used M_{age}=8.99; range=7-11, with reports from their main caregivers regarding 26 mental health problem symptoms taken from the Reporting Questionnaire for Children (RQC), Child Behavior Checklist (CBCL), and the Brief Screening and Diagnostic Questionnaire (CBTD). Classical test theory and factor analysis were conducted to analyze the classical location and information of each item, along with the dimensionality, reliability, and convergent validity of the scale. Item Response Theory (IRT) was used in order to estimate theoretically invariant item parameters for location and information. Findings reveal that the mental health scale for children has adequate psychometric properties for its use in Colombia. Furthermore, IRT analyses reveals a set of items that maximize information and that may be used in future administrations when more efficiency is warranted.

Keywords: Children's Mental Health, Classical Test Theory, Item Response Theory, Psychometrics, Test Information Function.

Cómo mejorar la medición de los problemas de salud mental en los niños colombianos mediante la Teoría de Respuesta al Ítem

Resumen

El estudio examina las propiedades psicométricas de la escala de salud mental para niños utilizada en la Encuesta Nacional de Salud Mental Colombia del 2015. Se utilizó una muestra representativa a nivel nacional de 2,727 niños M_{age}=8.99; rango=7-11, con informes proporcionados por sus cuidadores principales respecto de los síntomas de 26 problemas de salud mental tomados del Cuestionario de Reporte para Niños (RQC), el Inventario de Comportamiento de Niños (CBCL) y el Cuestionario Breve de Tamizaje y Diagnóstico (CBTD). Se emplearon la Teoría Clásica de los Tests y el análisis factorial para analizar la localización clásica y la información de cada ítem, así como la dimensionalidad, la confiabilidad y la validez convergente de la escala. Además, se utilizó la Teoría de Respuesta al Ítem (TRI) para calcular los parámetros de ítem teóricamente invariables para localización e información. Los resultados muestran que la escala de salud mental para niños tiene propiedades psicométricas adecuadas para su uso en Colombia. Además, los análisis TRI revelan un conjunto de ítems que maximizan la información y pueden ser usados en administraciones futuras en las que se requiera mayor eficiencia.

Palabras clave: función de información de la prueba, psicometría, salud mental de los niños, Teoría Clásica de los Tests, Teoría de Respuesta al Ítem.

Como melhorar a medição dos problemas de saúde mental nas crianças colombianas a partir da Teoria de Resposta ao Item

Resumo

Este estudo analisa as propriedades psicométricas da escala de saúde mental para crianças utilizada na Pesquisa Nacional de Saúde Mental Colômbiana de 2015. Foi utilizada uma amostra representativa no âmbito nacional de 2,727 crianças M_{age}=8.99; faixa etária=7-11, com informações fornecidas por seus cuidadores principais a respeito dos sintomas de 26 problemas de saúde mental tomados do *Reporting Questionnaire for Children*, do Inventário de Comportamentos de Crianças e Adolescentes (*Child Behavior Checklist*) e do Questionário Breve de Rastreamento e Diagnóstico. Foram utilizadas a Teoria Clássica dos Testes e a análise fatorial para analisar a localização clássica e a informação de cada item, bem como a dimensionalidade, a confiabilidade e a validade convergente da escala. Além disso, a Teoria de Resposta ao Item (TRI) para calcular o padrão de cada item teoricamente invariável para localização e informação. Os resultados indicam que a escala de saúde mental para crianças tem propriedades psicométricas adequadas para seu uso na Colômbia. Ainda, as análises com a TRI revelam um conjunto de itens que maximizam a informação e podem ser usados futuramente com mais eficácia.

Palavras-chave: função de informação do teste, psicometria, saúde mental das crianças, Teoria Clássica dos Testes, Teoria de Resposta ao Item.

RECENT ESTIMATES suggest that between 10 to 20 percent of children and adolescents in low- and- middle-income countries (LMICS) suffer from mental health problems (Erskine et al., 2017; Kieling et al., 2011). Despite this high prevalence, the mental health needs of children living in LMICS are often unattended for lack of funding, political indifference, or lack of qualified clinicians (Kieling et al., 2011). Children's mental health problems are particularly prevalent in conflict-affected countries, but these countries may, in turn, be less capable of monitoring their children's mental health needs (Dimitry, 2012). In this context, it is fundamental to develop efficient and reliable instruments to assess and monitor children's mental health to make the problem visible and to inform public policy efforts aimed at reducing it.

Colombia has suffered from more than 50 years of civil conflict, leaving more than 1.4 million children and adolescents as direct victims (Red Nacional de Información, 2018). Until 2015, the country did not have information about the national prevalence of mental health problems in children, despite evidence suggesting it was high in specific regions (OIM, UNICEF, & ICBF, 2013). In 2015, Colombia carried out its first nationally representative mental health survey for children aged seven to 11 years (ENSM, according to its acronym in Spanish; Ministerio de Salud & Colciencias, 2015), which included a 24-item scale to measure children's mental health symptoms. Subject-matter experts selected the 24 items included in the ENSM from three different existing scales that are briefly explained below (Rodriguez et al., 2016): the Reporting Questionnaire for Children (RQC; Giel et al., 1981), the Child Behavior Checklist (CBCL; Achenbach, 1999), and the Brief-Screening Diagnostic Questionnaire (CBTD according to its acronym in Spanish; Caraveo y Anduaga, 2007).

First, the RQC is a 10-item scale developed by the World Health Organization (WHO) to screen for significant degrees of emotional and behavioral disorder or psychotic disorders (Castro, Billick, & Swank, 2016). The target population for the RQC is children between the ages of five and 15. The main caregiver responded the questionnaire. Sample items include "Does the child wet or soil himself/herself?" and "Does the child tend to be alone?. The RQC has been administered in several countries, including Iraq, Ethiopia, Sudan, Philippines, India, and Colombia (Giel et al., 1981). Previous evidence shows that the RQC has a similar predictive power of children's clinical disorders to that of the CBCL in Iraqui Kurdistan children (Ahmad et al., 2007). Second, the CBTD is a 27-item questionnaire for parents, which comprises 10-items taken from the RQC and additional items reflecting additional mental health problems symptoms. The CBTD is intended to characterize common mental health problems, as well as hyperactivity, sadness, attention deficits, impulsivity, antisocial behavior (Caraveo and Anduaga, 2007). The CBTD has been widely used in Mexico (Caraveo and Anduaga, 2007).

Lastly, the CBCL is a parent-report questionnaire to detect emotional and behavioral problems in children and adolescents (Achenbach, 1999). The instrument targets children between the ages of six and 18 and consists of 113 Likert-scale items (i.e., absent, occurs sometimes, and occurs often) that assess the presence of different symptoms of mental health problems in the past six months. Besides, the 113-item, the CBCL has eight sub-scales, intended to measure anxiety, depression, somatic complaints, thought problems, attention problems, rule-breaking behavior, and aggressive behavior (Achenbach & Ruffle, 2000). Sample items include "can't concentrate, can't pay attention for long" and "complains of loneliness." The CBCL has been used in more than 30 countries, showing adequate psychometric properties in North America, and samples from Asia, Africa, South America, the Caribbean, and Europe (Ivanova et al., 2007).

Even though the content validity of the 24-items included in the ENSM scale was discussed (see Ministerio de Salud & Colciencias, 2015), the psychometric properties of the overall scale, which comprises items from the RQC, CBCL, and CBTD, have not been systematically assessed previously.

The present study seeks to fill this gap by analyzing the mental health problems' scale dimensionality, reliability, and convergent validity. Moreover, this study seeks to analyze evidence on the information each item provides, to define whether a more efficient scale (i.e., shorter and with a high level of information) is feasible, which might facilitate its future implementation.

Even though Classical Test Theory (CTT) is the most widely used framework to analyze the psychometric properties of test scores, Item Response Theory (IRT) offers unique features to improve the efficiency of a scale. Contrary to CTT, where item statistics (e.g., percentage of items correct, item-test correlation, measures of reliability) are population-dependent (Lord, Novick, & Birnbaum, 1968), the IRT estimates of item characteristics are assumed to be invariant across populations, occasions, and independent of other items embedded in the test or questionnaire (Brennan, National Council on Measurement in Education, & American Council on Education, 2006). In particular, IRT assesses item discrimination, which refers to the extent to which the item is capable of distinguishing between individuals with different levels of the latent trait, and item location. Item location refers, in this context, to the level of the latent trait where the scale is most reliable and precise in distinguishing between individuals (Embretson & Reise, 2013).

Another feature that makes IRT stand out is that it recognizes that a scale will be more reliable and precise at distinguishing between individuals at a certain segment of the latent continuum, whereas CTT assumes a single, homogenous estimate of reliability (Brennan et al., 2006). For this reason, IRT provides information that can be used for the design of most efficient scales, allowing the selection of items that provide more information at the levels of the latent trait of interest (Jessen, Ho, Corrales, Yueh, & Shin, 2018). In doing so, it is possible to obtain a shorter, easier to implement scales, as well as a set of items that allow a more reliable and targeted measure.

In the case of the mental health scale for children used in the ENSM, it is unclear whether the 26 items selected by subject-matter experts represent a single underlying dimension, hypothesized to be mental health problems. Moreover, it is unclear whether a future implementation, using a shorter but high-informing scale, is possible. These issues are critical not only to reduce the time and resources used at measuring children's mental health problems in Colombia in its post-conflict situation, but also to do so with precision, which is key to inform prevention and attention efforts across the country. This study contributes to these objectives by answering the following research questions:

1. Does the mental health scale used in the ENSM measure a single factor (i.e., mental health problems) as intended?
2. What are the psychometric properties of the mental health scale, according to CTT and IRT frameworks?
3. Is it possible to implement a more efficient (i.e., with fewer items and high precision) scale for children's mental health problems on a future occasion?

Methods

Participants

The ENSM is a nationally representative sample for non-institutionalized children aged seven to 11 years, representing four regions (Atlantic, Western, Central, and Pacific), Bogotá, and each of the 32 national departments. The ENSM sampling comprises a probabilistic, multistage sampling procedure, and the sample size was designed following findings from previous national studies (Rodríguez et al., 2016). The sample used in the present study includes 2,727 children, having complete information for all the cases included in the ENSM. According to the ENSM, the children were, on average, nine years old, and a little more than half of them were girls. In the sample, 19 percent of children belonged to an ethnic minority, around 58 percent lived with their parents, and 86 percent with their mother. Additionally, 98 percent attended school, and 21 percent were considered poor according to a multidimensional poverty index (Alkire & Foster, 2011). Table 1 presents details. ENSM surveys were collected between January and May 2015.

Table 1
Sample Characteristics (n=2,727)

Variable	M	SD	Min	Max
Age	8.99	1.41	7	11
Sex (=1 if male)	.49	.50	0	1
Ethnic minority	.19	.39	0	1
Lived with father	.58	.49	0	1
Lived with mother	.86	.31	0	1
Maternal education				
Less than basic	.05	.22	0	1
Basic	.27	.44	0	1
Secondary	.52	.49	0	1
Superior	.17	.37	0	1
Attended school	.98	.15	0	1
Reported-health status	4.02	.99	0	5
Multidimensional poor household	.21	.41	0	1
Region				
Central	.24	.42	0	1
Atlantic	.24	.42	0	1
Bogotá D.C.	.14	.35	0	1
Western	.21	.41	0	1
Pacific	.17	.37	0	1

Note: Averages using sample weights

Instruments

The ENSM included a 26-item instrument to assess children's mental health problems. This instrument includes 10 items from the RQC (Giel et al., 1981), and additional items from the CBCL (Achenbach, 1999), the CBTD (Caraveo and Anduaga, 2007), and "others based on the experience of the research groups [i.e., subject-matter experts that participated in the ENSM]" (Rodriguez et al., 2016, p. 15). The 26 items included in the scale, presented in Appendix 1, refer to *yes* (coded as 1) or *no* (coded as 0) questions, aimed at identifying diverse symptoms of mental health problems in children. Children's main caregivers filled out the questionnaire. In this case, 80 percent were their mothers, 7.2 percent their fathers, and the remaining were other caregivers. Even though previous studies show that the RQC, CBCL, and CBTD have good psychometric properties for assessing children's mental health problems (e.g., Ahmad et al., 2007; Castro et al., 2016), to date there is no validity evidence on the internal structure (i.e., coherence) of the scale employed in the Colombian ENSM.

Statistical Analysis

To begin with, Classical Test Theory (CTT) statistics were estimated (Crocker & Algina, 1986; Novick, 1966; Traub & Rowley, 1991). To analyze the characteristics of each item, the percentage of affirmative answers was used as a classical item location estimate and the item-test correlation as a classical information estimate. For the overall scale, Cronbach's alpha was estimated to examine the reliability (i.e., internal consistency) of the scale. Subsequently, factor analysis was used to fit a unidimensional model to the data and assess the dimensionality of the scale by analyzing the share of variance accounted for by the first factor (Merino-Soto, López-Fernández & Grimaldo-Muchotrigo, 2019; Thompson, 2004).

Furthermore, following the model presented in Equation 1, a two-parameter IRT model was fitted to the data (Embretson & Reise, 2013; Lord et al., 1968). In the model, θ_p represents the latent score (i.e., mental health problems–symptoms) of each child p, which is standardized (i.e., mean of zero and standard deviation of 1). Additionally, α_i represents the discrimination parameter, which indicates how well an item can distinguish between children with slightly different levels of the latent variable, and it is similar to factor loadings in a confirmatory factor analysis where items are continuous. Particularly, the discrimination parameter shows that a 1 unit increase in the latent variable θ produces an α increase in the log of the odds of answering the item affirmatively. Finally, b_i represents item location, which shows the level of the latent variable θ at which children have even odds of answering each item affirmatively. An advantage that IRT has over CTT is that the former estimates parameters that are invariant to populations of items and individuals, whereas the latter produces population-dependent parameters (Embretson & Reise, 2013; Traub & Rowley, 1991).

Equation 1

$$P_i(\theta_p) = \frac{1}{1 + \exp(-a_i(\theta_p - b_i))}\ ;\ \theta_p \sim N(0,1)$$

After fitting the IRT model to the data, the test information function is estimated as the sum of all item's information functions, which are calculated through their discrimination parameters (α), and the product of the likelihoods of having an affirmative (P) or negative (Q) answer in the item (Equation 2).

Equation 2

$$I(\theta) = \sum_i I_i(\theta_p) = \sum_i a_i^2 P_i(\theta) Q_i(\theta)$$

Using the test information function, the conditional errors of measurement (SE) were computed as presented in Equation 3. Contrary to CTT, where there is a single standard error of measurement for the scale, the (SE) in IRT shows the estimated error at different locations of the scale, making it possible to assess the level of imprecision in the measurement at different levels of θ.

Equation 3

$$SE(\hat{\theta}|\theta) = \frac{1}{\sqrt{I(\theta)}}$$

Subsequently, items that provided less information were flagged for potential exclusion in a shortened version of the scale. Particularly, the information provided by the scale comprising all the original items (i.e., 26) was compared with a reduced scale with 21 items, 18 items, 14 items, and 11 items, removing items that provided less information in a step-wise fashion. To provide validity evidence based on correlation for the test scores calculated with different items, convergent validity evidence was analyzed, using information gathered using the Diagnostic Interview for Children (DISC-IV; Shaffer, Fisher, Lucas, Dulcan, & Schwab-Stone, 2000), which is an instrument that evaluated 32 common psychiatric diagnoses of children based on the Diagnostic and Statistical Manual – IV (DSM-IV; Bell, 1994). Convergent validity evidence was also examined through correlations between the total score and children's reported physical health (Aarons et al., 2008), exposure to bullying or discrimination (Cooke, Bowie, & Carrère, 2014), and exposure to selected adversities (e.g., Anda et al., 2006). The latter includes exposure to factors such as community crime, parental separation, major sickness, and other stressful events, which have been widely linked to mental health problems during childhood. All analyses were conducted in Stata 15.1 (StataCorp, 2017).

Results

Classical Test Theory and Dimensionality

The 26-item scale exhibits good reliability, with a Cronbach's alpha (α) of .74, suggesting that 74 percent of observed score variance is accounted for by true score variance, according to CTT. As shown in Table 1, according to CTT estimates, all the items have high location parameters (i.e., only a small proportion of children present the assessed symptoms, see also Figure A1 for histograms), whereas there is considerable variability in the amount of information that each item provides (according to item-test correlation), ranging from .20 (item 21, "Has the child needed to change school more than 3 times?") to .58 (item 12, "Have you noticed that the child has difficulty making friends of his or her same age?"). Nonetheless, these parameters are population dependent, so in a different administration with a different sample, they may vary.

Table 2
Classical Test Theory Analysis (n=2,727; α=.74)

	Classical location	Classical information
	Affirmative (%)	Item-test correlation
Item 1	2.4	.23
Item 2	.9	.24
Item 3	4.2	.41
Item 4	3.7	.38
Item 5	1.8	.34
Item 6	1.8	.31
Item 7	12.4	.55
Item 8	18.6	.51
Item 9	6.8	.39
Item 10	4.8	.36
Item 11	5.0	.30
Item 12	15.4	.58
Item 13	3.0	.29
Item 14	3.7	.35
Item 15	7.9	.47
Item 16	8.1	.27
Item 17	7.9	.30
Item 18	9.0	.43
Item 19	2.7	.31
Item 20	1.3	.35
Item 21	.8	.20
Item 22	2.0	.23
Item 23	12.6	.47
Item 24	6.8	.54
Item 25	19.5	.27
Item 26	8.8	.34

Note: Table A1 presents items prompts.

The Kaiser-Meyer-Olkin of .81 shows that the sample is adequate for conducting factor analysis (Kaiser, 1974). A factor model was fit to determine whether a 1-factor solution could represent the data. As shown in Figure 1, the first factor explained 76 percent of the total variance, suggesting that the mental health scale used in the ENSM is capturing a single underlying dimension (i.e., mental health problems). As shown in Figure 2, a summary score following a single factor solution provides a skewed scale, as was expected, given the low prevalence of the different items assessed and given the purpose of the measurement (i.e., to identify mental health problems).

IRT and Scale Information

Table 3 summarizes item discrimination and location parameters for the 2pl-IRT model fitted to the data. Consistent with the findings from CTT, items have high location parameters, ranging from 1.31 (item 12, "Have you noticed that the child has difficulties making friends of his or her same age?") to 4.71 (item 25, "Is the child eating too little and

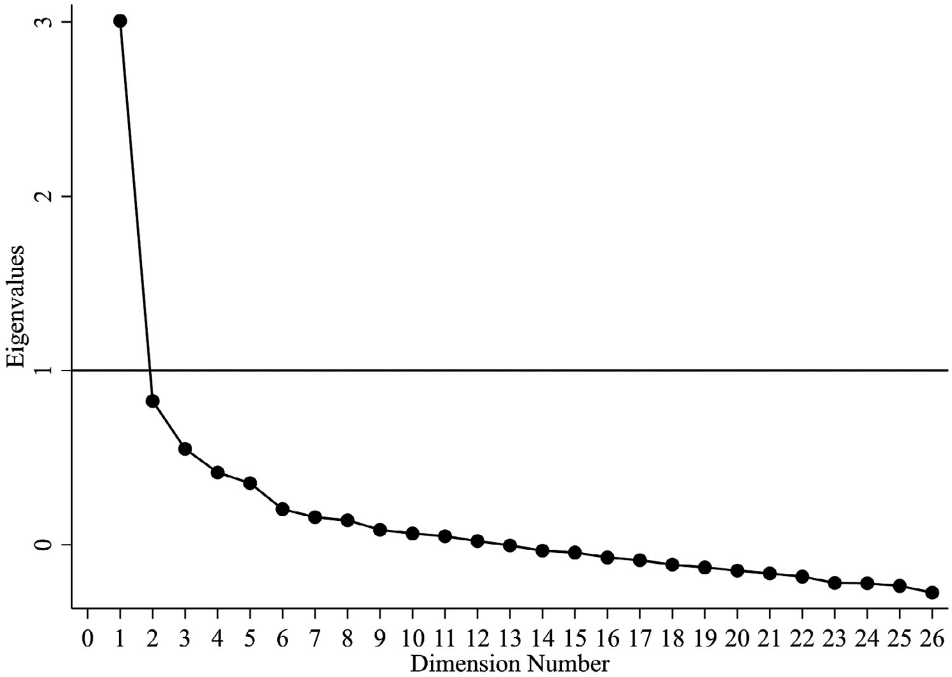


Figure 1. Scree plot of eigenvalues. This figure shows the variation accounted for by each dimension (out of 26) based on a factor analysis of standardized variables (n=2,727)

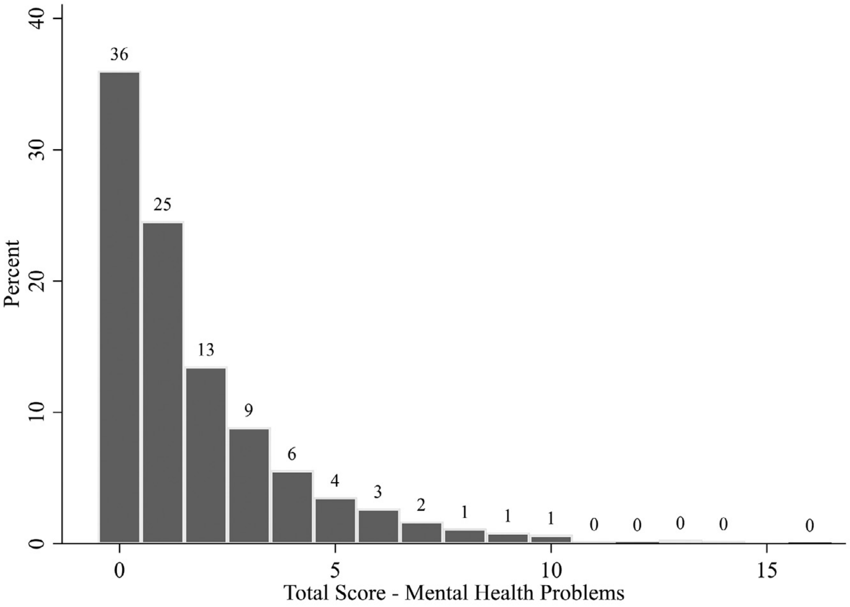


Figure 2. Scale total score – CTT score

losing weight?”) standard deviations above the mean. Furthermore, there is a larger dispersion in the estimated discrimination, ranging from 0.3 (item 25) to 2.54 (item 24, “Do you think the child is overeating for his age?”). Nonetheless, in general, most items have *a*’s parameters above one, suggesting that they distinguish among children with different levels of mental health problems. Figure A2 presents Item Characteristic Curves (ICC), where higher location parameters shift the ICC to the left, and a steepest slope reflects a higher discrimination parameter.

Even though high discrimination parameters are warranted to distinguish among children with different levels of mental health problems, the information provided by each item depends on its location along the latent scale. In the case of the mental health scale, most items have high location values, indicating that the items can distinguish among (and provide information for) children with the presence of mental health problems (which is the purpose of this measurement), as shown in Figure 3. Figure 4 also reflects this fact, showing that the overall scale provides more information for higher levels of theta. Consequently, the scale is more precise and reliable at higher levels of theta, having a lower conditional standard error of measurement between θ=2–3 SD above the mean.

Table 3
Item Discrimination and Location Parameters based on a 2pl-irt Model (n=2,727)

	Discrimination parameter estimates (*α*)	**Location parameter estimates *(b)***
Item 1	1.16	3.69
Item 2	1.96	3.21
Item 3	1.78	2.42
Item 4	1.71	2.57
Item 5	2.24	2.66
Item 6	1.93	2.85
Item 7	2.15	1.49
Item 8	1.46	1.37
Item 9	1.40	2.39
Item 10	1.50	2.57
Item 11	1.03	3.26
Item 12	2.20	1.31
Item 13	1.38	3.07
Item 14	1.66	2.60
Item 15	1.80	1.96
Item 16	.73	3.58
Item 17	.83	3.26
Item 18	1.31	2.22
Item 19	1.47	3.04
Item 20	.86	2.81
Item 21	1.63	3.63
Item 22	1.32	3.50
Item 23	1.30	1.89
Item 24	2.54	1.80
Item 25	.30	4.71
Item 26	.94	2.83

Note: Figure A2 presents item characteristic curves (ICC).

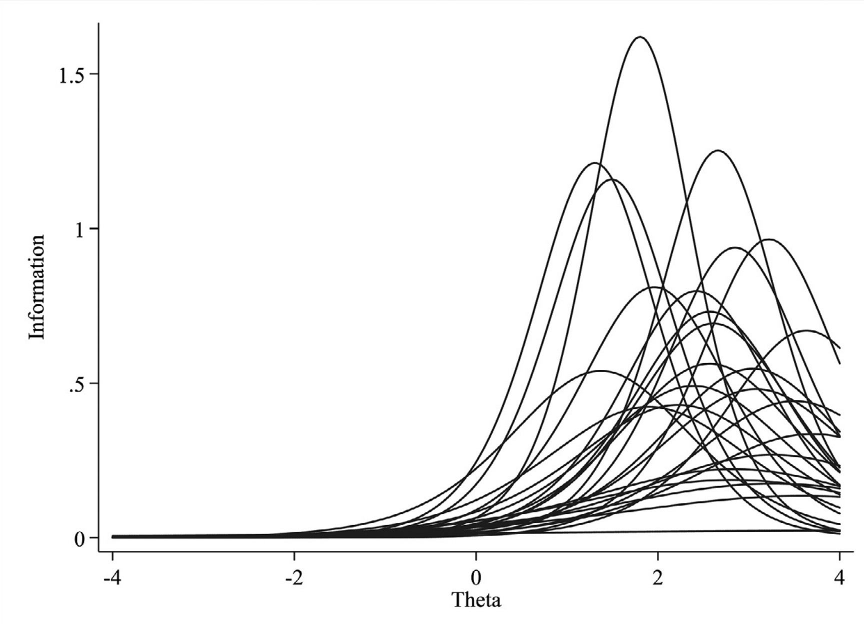


Figure 3. Item information functions (IIF) from a 2pl-IRT model (N=2,727).

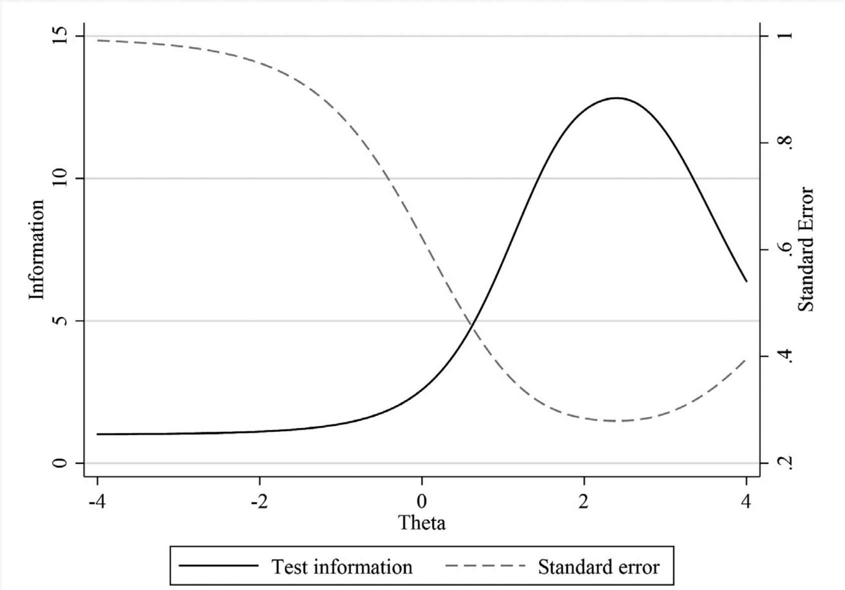


Figure 4. Test information function and conditional standard error of measurement from a 2pl-IRT model (N=2,727).

IRT's main assumption is local independence, which indicates that theta (i.e., the level of the latent trait) provides all the information needed to know the probability of an affirmative response to an item. Given this assumption, it is possible to estimate the information provided by different scales by adding or subtracting the corresponding item information functions (Jessen et al., 2018). Figure 5 presents the test information functions for scales composed of all the 26 items, 21 items, 18 items, 14 items, and 11 items, subtracting items with lower levels of information in a stepwise fashion (see Table A2 for additional details). The reduction of items produces lower levels of information, but the reduction is not considerable. In general, even using the 11 most informing items would produce a reliable measure for high levels of theta (particularly around 2 to 3 SD above the mean).

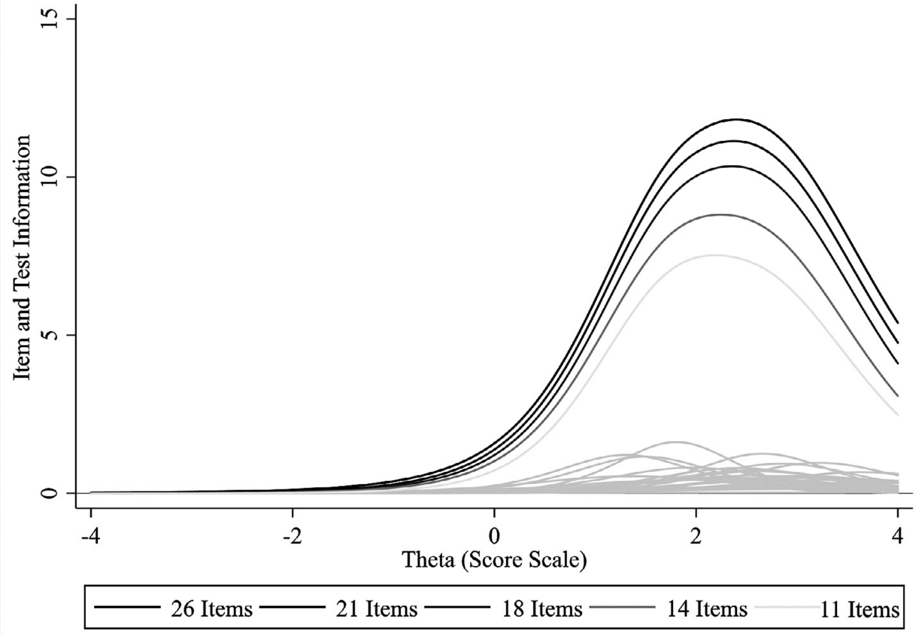


Figure 5. Test information functions from a 2pl-IRT model for different scale specifications (N=2,727).

Table 4
Correlation Coefficients

	Convergent Validity			
	DISC-IV Problems	Physical health	Discriminated	Not exposed to adversities
26-item scale (α=.74)	.25**	-.29**	.34**	-.19**
21-item scale (α=.74)	.25**	-.27**	.34**	-.18**
18-item scale (α=.73)	.24**	-.26**	.34**	-.17**
14-item scale (α=.70)	.22**	-.25**	.29**	-.15**
11-item scale (α=.66)	.23**	-.25**	.29**	-.15**

*Note:*** *p*<.01

Lastly, the correlational analysis reveals that the mental health scale has a statistically significant association with the expected sign to the number of psychiatric disorders identified using the DISC-IV, as well as with reported physical health, being discriminated, and not being exposed to adversities (Table 4). All these correlations keep their significance and have similar magnitudes when using reduced forms of the scale, suggesting that a shorter scale may be used if needed, given that it is reliable (Figure 5) and has correlational validity evidence (Table 4).

Discussion

In 2015, Colombia undertook its first nationally representative mental health survey (i.e., the ENSM) for children aged seven to 11 years. The ENSM included 26 items that were hypothesized to be measuring children's mental health problems, taken from the RQC (Giel et al., 1981), CBCL (Achenbach, 1999), CBTD (Caraveo and Anduaga, 2007), and others were based on the expertise of the group of researchers (Rodriguez et al., 2016). The items were based on measures whose score interpretation has validity evidence, offering content validity for the ENSM scale. Nonetheless, little was known about validity evidence based on coherence (i.e., internal structure). Moreover, it was not known whether a shorter scale would provide similar levels of information, thus being more efficient while having a high level of measurement precision.

The purpose of this study was to analyze the children's mental health scale using the CTT and IRT frameworks. The findings indicate that the scale has adequate internal consistency reliability, and the evidence from factor analysis suggests it is measuring a single latent construct. Furthermore, results from an IRT model show that most items have a high location as can be expected, reflecting the fact that only individuals with a high θ (i.e., exhibiting mental health problems) will have even or higher odds of answering each item affirmatively. Given the local independence assumption, IRT also reveals that different items provide substantial different levels of information to the total scale information, suggesting that a more efficient scale, employing only high-informative items, would be feasible. Indeed, findings from the item information function and convergent validity indicate that shorter scales will keep desirable psychometric properties and could be employed in future implementations of the mental health scale when increased efficiency is needed.

One major strength of this study and contribution to the Colombian literature is the use of an IRT framework, which conversely to CTT theoretically offers population-invariant parameters that can accurately inform future implementations of each item. Indeed, the CTT framework produces parameters that depend on the specific population where the items are implemented, and which are, to a certain extent, predictable. For instance, following the Spearman-Brown prophecy formula, it is possible to infer that Cronbach's alpha will be higher as one test has more items and as the population where the test is implemented is more heterogeneous, whereas fewer items and a more homogenous population would lead to lower test reliabilities (Traub & Rowley, 1991). On the other hand, IRT estimates item location, discrimination, and information that are assumed to hold in different occasions and populations (Embretson & Reise, 2013). These population-invariant parameters can inform the design of scales, maximizing precision at the desired θ level and permitting the selection of high-informing items when efficiency is paramount (Jessen et al., 2018).

Even though this study makes relevant contributions offering validity evidence based on coherence and correlation for the ENSM children's mental health scale, it does not provide validity evidence based on response process (i.e., cognition) or consequences (Koretz, 2008). A future pilot study could be implemented to analyze the type of cognitive process respondents employ to respond to the scale's items, assessing whether some items may be particularly cognitive-demanding. Moreover, it could be useful to conduct studies that make it possible to elucidate whether the test produces certain consequences on respondents, such as changes in their behaviors or interactions with their children following the test. It is also important to consider that the ENSM scale faces limitations that future studies should explore further. For example, the scale is based on parental reports and these reports may be biased due to parents' mental health problems. Future efforts must be conducted to offer more evidence on the predictive validity of the scale on children's mental health problems in Colombia according to clinical assessments.

Conclusion

The 26-item children's mental health scale used in the ENSM has adequate psychometric properties, and evidence from factor analysis suggests it is measuring a single latent construct. A 2pl-IRT model reveals that the scale is accurate at distinguishing between children with high levels of θ, around one and three SD above the mean. In a future implementation of the scale, when lowering the number of items and higher efficiency are needed, a 21, 18, 14, and even 11- item scale may hold desirable properties and predictive power. These findings suggest that future efforts can be conducted to continue monitoring children's mental health in Colombia, especially in the post-conflict situation, when it is necessary to identify children who would need additional supports.

References

Aarons, G. A., Monn, A. R., Leslie, L. K., Garland, A. F., Lugo, L., Hough, R. L., & Brown, S. A. (2008). Association between mental and physical health problems in high-risk adolescents: a longitudinal study. *The Journal of Adolescent Health : official Publication of the Society for Adolescent Medicine, 43,* 260-267. https://doi.org/10.1016/j.jadohealth.2008.01.013

Achenbach, T. M. (1999). The Child Behavior Checklist and related instruments. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcomes assessment* (pp. 429-466). NJ, US: Lawrence Erlbaum Associates Publishers.

Achenbach, T. M., & Ruffle, T. M. (2000). The Child Behavior Checklist and Related Forms for Assessing Behavioral/Emotional Problems and Competencies. *Pediatrics in Review, 21,* 265-271. https://doi.org/10.1542/pir.21-8-265

Ahmad, A., Abdul-Majeed, A. M., Siddiq, A. A., Jabar, F., Qahar, J., Rasheed, J., & von Knorring, A.-L. (2007). Reporting Questionnaire for Children as a Screening Instrument for Child Mental Health Problems in Iraqi Kurdistan. *Transcultural Psychiatry, 44,* 5-26. https://doi.org/10.1177/1363461507074949

Alkire, S., & Foster, J. (2011). Counting and multidimensional poverty measurement. *Journal of Public Economics, 95,* 476-487. https://doi.org/10.1016/j.jpubeco.2010.11.006

Anda, R. F., Felitti, V. J., Bremner, J. D., Walker, J. D., Whitfield, C., Perry, B. D., . . . Giles, W. H. (2006). The enduring effects of abuse and related adverse experiences in childhood. *European Archives of Psychiatry and Clinical Neuroscience, 256,* 174-186. https://doi.org/10.1007/s00406-005-0624-4

Bell, C. C. (1994). DSM-IV: Diagnostic and statistical manual of mental disorders. *JAMA, 272,* 828-829. https://doi.org/10.1001/jama.1994.03520100096046

Brennan, R. L., National Council on Measurement in Education, & American Council on Education. (2006). *Educational measurement* (4th ed. ed.). Westport, US.

Caraveo-Anduaga, J. J. (2007). Validez del Cuestionario Breve de Tamizaje y Diagnóstico (CBTD) para niños y adolescentes en escenarios clínicos. *Salud Mental, 30,* 42-49. Retrieved from http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-33252007000200042&nrm=iso

Caraveo y Anduaga, J. J. (2007). Cuestionario breve de tamizaje y diagnóstico de problemas de salud mental en niños y adolescentes: algoritmos para síndromes y su prevalencia en la ciudad de México. Segunda parte. *Salud Mental, 30,* 48-55.

Castro, J., Billick, S. B., & Swank, A. C. (2016). Utility of a New Spanish RQC and PSC in Screening with CBCL Validation. *Psychiatric Quarterly, 87,* 343-353. https://doi.org/10.1007/s11126-015-9391-1

Cooke, C. L., Bowie, B. H., & Carrère, S. (2014). Perceived Discrimination and Children's Mental Health Symptoms. *Advances in Nursing Science, 37,* 299-314. https://doi.org/10.1097/ans.0000000000000047

Crocker, L. M., & Algina, J. (1986). Introduction to classical and modern test theory. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, US. 32887.

Dimitry, L. (2012). A systematic review on the mental health of children and adolescents in areas of armed conflict in the Middle East. *Child: Care, Health and Development, 38,* 153-161. https://doi.org/10.1111/j.1365-2214.2011.01246.x

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*: Psychology Press.

Erskine, H. E., Baxter, A. J., Patton, G., Moffitt, T. E., Patel, V., Whiteford, H. A., & Scott, J. G. (2017). The global coverage of prevalence data for mental disorders in children and adolescents. *Epidemiology and Psychiatric Sciences, 26,* 395-402. https://doi.org/10.1017/S2045796015001158

Giel, R., de Arango, M. V., Climent, C. E., Harding, T. W., Ibrahim, H. H. A., Ladrido-Ignacio, L., . . . Younis, V. O. A. (1981). Childhood Mental Disorders in Primary Health Care: Results of Observations in Four Developing Countries. *Pediatrics, 68,* 677-683.

Ivanova, M. Y., Achenbach, T. M., Dumenci, L., Rescorla, L. A., Almqvist, F., Weintraub, S., . . . Verhulst, F. C. (2007). Testing the 8-Syndrome Structure of the Child Behavior Checklist in 30 Societies. *Journal of Clinical Child & Adolescent Psychology, 36,* 405-417. https://doi.org/10.1080/15374410701444363

Jessen, A., Ho, A. D., Corrales, C. E., Yueh, B., & Shin, J. J. (2018). Improving Measurement Efficiency of the Inner EAR Scale with Item Response Theory. *Otolaryngology–Head and Neck Surgery, 158,* 1093-1100. https://doi.org/10.1177/0194599818760528

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39,* 31-36. https://doi.org/10.1007/bf02291575

Kieling, C., Baker-Henningham, H., Belfer, M., Conti, G., Ertem, I., Omigbodun, O., . . . Rahman, A. (2011). Child and adolescent mental health worldwide: evidence for action. *The Lancet, 378,* 1515-1525. https://doi.org/10.1016/S0140-6736(11)60827-1

Koretz, D. M. (2008). *Measuring up : what educational testing really tells us*. Cambridge, Mass.

Lord, F. M., Novick, M. R., & Birnbaum, A. (1968). *Statistical theories of mental test scores*. Oxford, England: Addison-Wesley.

Ministerio de Salud, & Colciencias. (2015) Encuesta nacional de salud mental 2015. In Bogotá, COL: Ministerio de Salud & Colciencias.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3,* 1-18. https://doi.org/10.1016/0022-2496(66)90002-2

OIM, UNICEF, & ICBF. (2013). *Estado psicosocial de los niños, niñas y adolescentes: una investigación de consecuencias, impactos y afectaciones por hecho victimizante con enfoque diferencial en el contexto del conflicto armado colombiano*. Bogotá, COL.

Red Nacional de Información. (2018). Registro Único de Victimas (RUV). Retrieved from https://www.unidadvictimas.gov.co/es/registro-unico-de-victimas-ruv/37394

Rodríguez, N., Rodríguez, V. A., Ramírez, E., Cediel, S., Gil, F., & Rondón, M. A. (2016). Aspectos metodológicos del diseño de muestra para la Encuesta Nacional de Salud Mental 2015. *Revista Colombiana de Psiquiatría, 45,* 26-30. https://doi.org/10.1016/j.rcp.2016.08.009

Rodriguez, V., Moreno, S., Camacho, J., Gómez-Restrepo, C., de Santacruz, C., Rodriguez, M. N., & Tamayo Martínez, N. (2016). Diseño e implementación de los instrumentos de recolección de la Encuesta Nacional de Salud Mental Colombia 2015. *Revista Colombiana de Psiquiatría, 45,* 9-18. https://doi.org/10.1016/j.rcp.2016.10.001

Shaffer, D., Fisher, P., Lucas, C. P., Dulcan, M. K., & Schwab-Stone, M. E. (2000). NIMH Diagnostic Interview Schedule for Children Version IV (NIMH DISC-IV): Description, Differences From Previous Versions, and Reliability of Some Common Diagnoses. *Journal of the American Academy of Child & Adolescent Psychiatry, 39,* 28-38. https://doi.org/10.1097/00004583-200001000-00014

StataCorp. (2017). Stata statistical software: release 15. College Station, TX, US: StataCorp LLC.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC, US: American Psychological Association.

Traub, R. E., & Rowley, G. L. (1991). Understanding Reliability. *Educational Measurement: Issues and Practice, 10,* 37-45. https://doi.org/10.1111/j.1745-3992.1991.tb00183.x

Appendix

Table A1. *Item prompts*

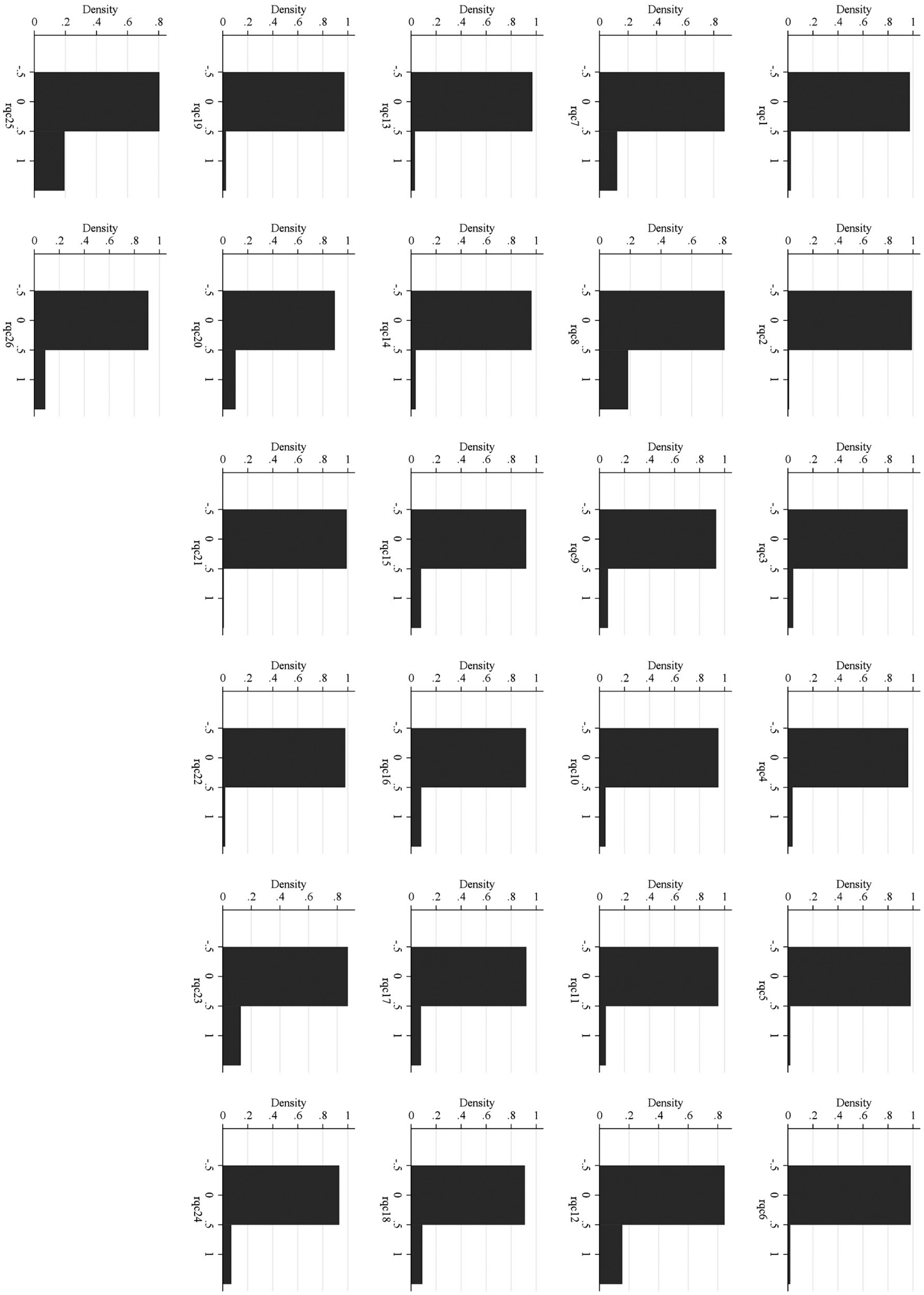
Item	Label	Label in original language
Item 1	Is the child speech in any way abnormal?	¿El lenguaje del niño es anormal en alguna forma?
Item 2	Does the child sleep badly?	¿El nino duerme mal?
Item 3	Did the child ever have a fit or fall to the ground for no reason?	¿Ha tenido el niño en algunas ocasiones convulsiones o caídas al suelo sin razón?
Item 4	Does the child suffer from frequent headaches?	¿Sufre el niño de dolores fuertes de cabeza?
Item 5	Does the child run away from home frequently?	¿El niño ha huido de la casa frecuentemente?
Item 6	Dies the child steal things from home?	¿Ha robado cosas de la casa?
Item 7	Does the child get scared or nervous for no good reason?	¿Se asusta o pone nervioso sin razón?
Item 8	Does the child in any way appear backward or slow to learn as compared with other children of about the same age?	¿Parece como retardado o lento para aprender?
Item 9	Does the child nearly never play with other children?	¿El niño casi nunca juega con otros niños?
Item 10	Does the child wet or soil himself/herself?	¿El niño se orina o defeca en la ropa?
Item 11	Has the child stopped talking seasonally or at all?	¿El niño ha dejado de hablar por temporadas o del todo?
Item 12	Have you noticed that the child has difficulty making friends of the same age?	¿Ha notado que al niño se le dificulte hacer amigos de su misma edad?
Item 13	Does the child tend to be alone?	¿El niño tiende a permanecer sólo?
Item 14	Does the child frequently walk with difficulty or accidentally hit himself?	¿El niño frecuentemente camina con dificultad o se golpea accidentalmente?
Item 15	Does the child exhibit strange behaviors such as talking alone without playing?	¿El niño presenta comportamientos extraños como hablar sólo sin estar jugando?
Item 16	Has the child had trouble learning to read or write?	¿el niño ha tenido problemas para aprender a leer o escribir?
Item 17	Has the child had trouble learning math?	¿El niño ha tenido problemas para aprender matemáticas?
Item 18	Has the child repeatedly been a victim of abuse or physical or psychological maltreatment?	¿El niño, repetidamente, ha sido víctima de abuso o maltrato, físico o psicológico?
Item 19	Has the child molested or repeatedly assaulted other children?	¿El niño ha molestado o ha agredido repetidamente a otros niños?
Item 20	Can't concentrate, can't pay attention for long?	¿El niño ha tenido problemas para fijar y mantener la atención o concentrarse?
Item 21	Has the child needed to change school more than 3 times?	¿El niño ha necesitado cambio de institución escolar más de 3 veces?
Item 22	Does the child refuse to go to school repeatedly?	¿El niño se niega a ir a la escuela repetidamente?
Item 23	Does the child have difficulty following rules, limits or respecting authority figures?	¿El niño tiene dificultad para seguir normas, límites o respetar figuras de autoridad?
Item 24	Do you think the child is overeating for his age?	¿Considera que el niño está comiendo en exceso para su edad?
Item 25	Is the child eating too little and losing weight?	¿El niño está comiendo muy poco y ha bajado de peso?
Item 26	Does the child repeatedly complain of pain, dizziness, desire to vomit or other ailments without medical explanation?	¿El niño se queja repetidamente de dolores de estómago, extremidades, de mareos, ganas de vomitar u otras dolencias sin explicación médica?

Table A2. *Different scale-specifications*

	Original scale	21-items	18-items	14-items	11-items
Item 1	X	X			
Item 2	X	X	X	X	X
Item 3	X	X	X	X	X
Item 4	X	X	X	X	X
Item 5	X	X	X	X	X
Item 6	X	X	X	X	X
Item 7	X	X	X	X	X
Item 8	X	X	X		
Item 9	X	X	X		
Item 10	X	X	X	X	
Item 11	X				
Item 12	X	X	X	X	X
Item 13	X	X	X		
Item 14	X	X	X	X	X
Item 15	X	X	X	X	X
Item 16	X				
Item 17	X				
Item 18	X	X			
Item 19	X	X	X	X	
Item 20	X	X	X	X	X
Item 21	X	X	X	X	
Item 22	X	X	X		
Item 23	X	X			
Item 24	X	X	X	X	X
Item 25	X				
Item 26	X				

Table A1 presents items prompts

Figure A1. Item histograms



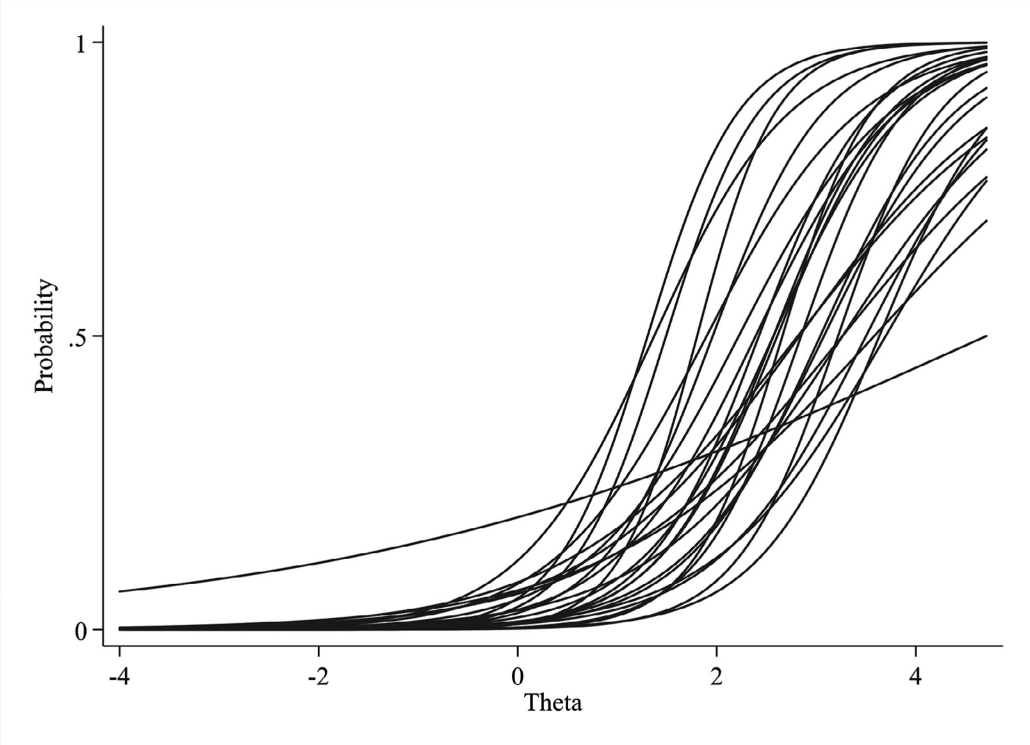


Figure A2. Item characteristic curves (ICC)